# Machine Learning-Assisted Tools for User eXperience Evaluation: A Systematic Mapping Study

Tales Schifelbein Soares
Universidade Federal do Pampa - Campus Alegrete
Alegrete, Rio Grande do Sul, Brazil
talessoares.aluno@unipampa.edu.br

Ricardo Luiz Hentges Costa
Universidade Federal do Pampa - Campus Alegrete
Alegrete, Rio Grande do Sul, Brazil
ricardohentges.aluno@unipampa.edu.br

Estefano Soares
Universidade Federal do Pampa - Campus Alegrete
Alegrete, Rio Grande do Sul, Brazil
estefanosoares.aluno@unipampa.edu.br

Ivanilse Calderon
Instituto Federal de Educação, Ciência e Tecnologia de Rondônia (IFRO)
Porto Velho, Rondônia, Brasil
ivanilse.calderon@ifro.edu.br

Gabriel Machado Lunardi
Universidade Federal de Santa Maria
Santa Maria, Rio Grande do Sul, Brasil
gabriel.lunardi@ufsm.br

Pedro Henrique Dias Valle
Universidade de São Paulo
São Paulo, São Paulo, Brasil
pedrohenriquevalle@usp.br

Gilleanes T. A. Guedes
Universidade Federal do Pampa
PPGES - Campus Alegrete
Alegrete, Rio Grande do Sul, Brasil
gilleanesguedes@unipampa.edu.br

Williamson Silva
Universidade Federal do Pampa
PPGES - Campus Alegrete
Alegrete, Rio Grande do Sul, Brasil
williamsonsilva@unipampa.edu.br

## ABSTRACT

**Context**: Information Systems (IS) have grown exponentially, significantly influencing professional and personal environments. Both scenarios require a distinguished User Experience (UX), which generates positive feelings such as loyalty, learning, and satisfaction from end users. Consequently, tools, software, and applications that integrate Machine Learning (ML) techniques with UX are necessary for enhancing the quality of IS and increasing the productivity of UX specialists. **Problem**: There is a continued need for more experimental evidence regarding the development, employability/applicability, evaluation, and evolution of current technologies that automate manual tasks performed by experts. Specifically, such technologies aim to reduce workload, eliminate evaluation biases, and identify patterns that might go unnoticed during assessments. **Method**: This work aims to summarize and characterize, through a Systematic Mapping Study (SMS), the tools that employ ML techniques to assist in the UX evaluation process. To help us, we defined seven sub-questions that will be addressed based on the data collected from the selected studies. **Contributions and Impact**: Based on the selected studies, we analyzed and characterized the assessment tools to provide a comprehensive understanding for both the academic and professional communities. This work presents the current state of tools that integrate ML techniques for UX evaluation, offering valuable insights into their effectiveness and application within the IS domain.

## KEYWORDS

User Experience, Machine Learning, Systematic Mapping Study

## 1 INTRODUCTION

Information Systems (IS) has increased exponentially, significantly influencing professional and personal environments [35, 54]. Companies adopt IS to optimize processes [39], while end users use it for shopping and communication purposes [40] or entertainment. This growth generates a demand for interfaces with a good experience, generating positive feelings such as loyalty, learning, and user satisfaction [68].

In this sense, the User Experience (UX) studies ways of measuring user satisfaction and identifying ways to improve it, aiming at increasing immersion and satisfaction, building a good relationship between the end user and the IS [68]. Positive experiences influence end-user satisfaction and loyalty [50, 68], thus vital to the success and acceptance of SI by end-users [29]. On the other hand, negative experiences can lead to the abandonment of [50], besides discouraging other potential users [43].

With this, software development companies are encouraged to design IS with high quality (process and product), aiming to promote unique and positive experiences for end users and obtain a competitive advantage before the software industry [12]. In particular, UX studies have become increasingly important and necessary to support software engineers and systems analysts during the design and SI evaluation [58, 66]. Although industry and academia are interested in UX, designing, evaluating, and measuring UX and identifying which factors most impact the user experience is a very complex task [9, 48].

UX evaluations performed by professionals in the area are subjective by nature [30] because each evaluator has their perceptions, preferences, and previous experiences, which can generate divergences between them [8]. In addition, UX assessments directly depend on self-reported measurements and/or observations collected by researchers who may be unable to discover the true emotional experience of the user [14]. The lack of systematic analysis in usability assessments can lead to neglecting important problems, including those that could be considered unsatisfactory experiences according to established standards. [52]. These evaluations require

significant investments of time and financial resources [11, 27], besides presenting a high degree of difficulty since they involve tests with real users, interviews, and detailed analysis of the interactions that can be collected through logs, videos, and audio and relate this data collection to various extensions of the UX [27]. Despite this, the literature presents several initiatives that have been proposed to minimize these challenges [11, 14, 27, 30].

In this scenario, there was an increase in the number of initiatives proposed in the scientific literature to evaluate UX at different stages of software development [66]. In particular, Rivero and Conte [58], through an extension of the work of Vermeeren et al. [66], presented an overview of the types of UX evaluation methods identified and classified them. However, there is little experimental evidence on the development, employability/applicability, evaluation, and evolution of current technologies that enable UX specialists to automate manual tasks and require much time from the researchers involved. Therefore, there is a need for tools that can assist professionals in reducing the workload, eliminating biases, and identifying patterns that could be easily ignored, in addition to optimizing the time of the experts to focus on other aspects of the assessment [52].

Machine Learning (ML) techniques are great allies in automation, recognized for their ability to identify patterns in different contexts when correctly applied [21]. Tools with integration between UX and ML can provide more optimized and consistent feedback with the pre-established standards, assisting UX professionals in data collection, either from websites or interviews during the process, as in the analysis process. In addition, these tools can adjust the interface, making it more intuitive and efficient, providing a good experience for the end user [9, 21] and providing autonomy and agility to the evaluator. The use of ML to recognize UX patterns in IS is a viable alternative and can offer important support to professionals in the area. Costa et al. [13] also reported that UX professionals can use ML to assist in UX analysis, increasing their ability to manage large and complex data sets and maximizing the time spent on each analysis.

This research aims to summarize and characterize, through a Systematic Mapping Study (SMS), the tools that employ ML techniques to evaluate UX. SMS categorizes and summarizes existing information on a research question unbiasedly [33]. This type of study helps identify gaps in current research to suggest areas for further investigation [33]. Thus, this SMS presents an overview of the current scenario and broadens researchers' understanding of this field. In addition, we expected to enrich the academic and professional IS community's understanding of ML's transformative role in UX research.

## 2 THEORETICAL BACKGROUND

### 2.1 *Machine Learning*

Machine Learning (ML), a sub-field of Artificial Intelligence (AI), has as its primary objective to learn through past experiences, autonomously creating a hypothesis or function from examples that can solve the problem in question [60]. Among the various existing ML tasks, grouping and classification stand out. The grouping seeks to find data groups with similar characteristics or properties, identifying similarities in datasets. The classification already associates data to classes, such as feelings, weights, or other classes relevant to the study [21]. The classification algorithms are evaluated by commonly derived metrics from a confusion matrix [21]. In specific areas, such as text mining, precision, recall, and f-measurement, ROC curves are used, among others [10].

Text Mining, as an application of ML in Natural Language Processing (NLP), seeks to extract patterns from large volumes of textual data [44]. A prominent area within Text Mining is Sentiment Analysis, which aims to identify a text's polarity and/or emotion. For example, social media analytics can adopt Sentiment Analysis to evaluate public opinion about a product, service, or event, providing decision-making possibilities for companies and organizations. ML tasks, such as classification, are used to train models that recognize linguistic patterns associated with different feelings, allowing one to automate the analysis of large amounts of textual data and extract relevant information about the opinions and emotions expressed [10].

The text pre-processing is a mandatory step for both the classification in Sentiment Analysis and the grouping task, although with some nuances [46]. In both functions, removing noise such as punctuation, special characters, and numbers is common and converting to lowercase to standardize texts. The removal of stop words - frequent words with little semantic meaning (e.g., 'e,' 'a', 'o') - is often applied but requires care, especially in clustering, in which these words may, in some contexts, contribute to the formation of relevant groups. Techniques such as lemmatization or stemming reduce words to their roots, help group variations of the same word, and simplify the textual representation. Tokenization, which divides the text into smaller units (words or phrases), is the basis for both tasks. Vectorization, a fundamental step to represent the text numerically, allows ML algorithms to process the information [45]. Methods such as Bag-of-Words (BoW), TF-IDF, and word embeddings are used, with word embeddings being particularly useful for capturing semantic relationships, which is important for Sentiment Analysis [**?** ]. The choice of the vectorization method depends on the characteristics of the data and the purpose of the analysis. Finally, in the classification for sentiment analysis, the vetored data is divided into training sets and tests to train and evaluate the model, which will teach how to associate the patterns in the vectors with the sentiment labels. In the cluster, the choice of pre-processing and vectorization techniques directly impacts the quality of the groups formed [46].

### 2.2 *User eXperience*

User Experience (UX) studies users' perceptions and responses to using products, systems, or services. According to the definition of ISO 9241-210 [18], UX involves the perceptions and responses resulting from using or anticipating a product. In particular, Hassenzahl [26] highlights that UX evaluates software quality from two main perspectives: (i) pragmatic (focused on effectiveness and efficiency of use) and (ii) hedonic (related to the stimuli and feelings of users during interaction). Thus, UX goes beyond the limits of usability, focusing only on pragmatic aspects, with usability being a subset of UX [26]. It is worth mentioning that, in recent years, there has been an increase in the number of methods proposed in the literature to evaluate UX at different stages of development [66]. Rivero

and Conte [58] provides an overview of these UX evaluation methods and categorizes them; however, there is still little experimental evidence on the applicability of (semi)automated UX methods, especially those supported by ML techniques, which constitutes the research gap explored in this work.

In a complementary way, the analysis of feelings, in turn, emerged in the early 2000s in the field of Natural Language Processing (NLP) to identify feelings in customer reviews and separate affective texts from factual ones. This field was made possible by increased digitized texts in the 1990s [49]. Several techniques are used to carry out this analysis, such as ML, lexical dictionaries, and PLN, in addition to the combination of these strategies [57]. However, many studies and datasets are still produced in English, requiring efforts to build databases in other languages, such as Portuguese. In analyzing feelings, the methods for identifying emotions are usually classification algorithms, such as Naive Bayes, which, although simple, is widely effective [59]. Another relevant algorithm is BERT, a deep learning approach developed by Google, which has shown positive results in feeling rating tasks [17]. Techniques such as lexical dictionaries and crowdsourcing are also used [57].

## 3 RELATED WORKS

Obaidi et al. [53] conducted a systematic mapping study (SMS) on applying sentiment analysis tools in Software Engineering. The authors analyzed 106 primary studies, examining the domains of application, purposes, data sets, development approaches, and the challenges faced in implementing these tools. The results show that most tools use neural networks. BERT stood out as the most effective in precision and F1-measurement due to its ability to understand the context of words. However, the study also identified significant challenges, such as difficulty dealing with irony and sarcasm, subjectivity in data labeling, and performance degradation in multi-platform scenarios, indicating that there is still room for improvements and innovations in analyzing feelings applied to Software Engineering.

The study by Abbas et al. [1] presented a Systematic Literature Review (SLR) on applying ML solutions to improve product usability. The research explored the challenges UX designers face when integrating ML into their design processes, identifying obstacles such as inadequate tools and limited technical knowledge about ML. The results highlight the need for more initiatives and tools to empower designers and emphasize the untapped potential of ML for personalization and innovation in UX. The study suggests that to maximize the use of ML, designers must understand both its technical aspects and the ethical implications of its application in the practice of design.

Virvou [67] conducted a critical analysis of scientific studies and published research on the interdependence between AI and UX. The authors comment that the relationship between AI and UX is characterized by reciprocity, where the quality of UX affects the effectiveness of AI, and the capabilities of AI, in turn, transform UX. The study reviews and critically analyzes previous research to understand this relationship, focusing on the contributions and limitations of AI in UX and vice versa, as well as the evolution of techniques such as machine learning and NPL. The study highlights best practices for the design of transparent, explainable, and fair AI systems, as well as proposing areas for future research, emphasizing the importance of taking into account the findings of Human-Computer Interaction (HCI) when designing AI-based systems, ensuring more excellent reliability, usability, and user control.

Lu et al. [41] conducted a Systematic Literature Review on applying AI tools to support UX design, focusing on user-centered approaches. The results highlight that simplified automation can interfere with empathy-building processes, which is essential for UX designers. The results also indicate that most AI research for UX follows a technology-driven approach rather than a user-centered perspective, suggesting the need for a better understanding of UX methodologies and objectives. The study concludes that there is a demand for metrics and data sets that align with the designer's mindset and for closer collaboration between the HCI and AI communities to provide more effective support.

This study differentiates itself from previous secondary studies by conducting an SMS exclusively focused on tools that use ML in UX analysis. While research such as Obaidi et al. [53] and Abbas et al. [1] address different aspects of ML applications in contexts like sentiment analysis and usability, the focus of this work is to identify and classify tools that employ ML specifically to automate and/or optimize UX evaluation. Unlike studies that explore general challenges and implementations, this research aims to consolidate knowledge about the use of ML in supporting UX analysis, highlighting how these tools assist, which algorithms they use, their domains, and how data is presented.

## 4 RESEARCH METHOD

The purpose of the SMS is to characterize knowledge about what is available to subsidize the conduct of UX assessments through ML techniques. In this sense, we conducted an SMS regarding the guidelines provided by Petersen et al. [55]. Figure 1 gives an overview of the steps undertaken in these SMS.

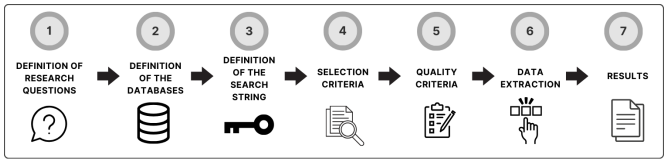



**Figure 1: SMS Steps**

### 4.1 Research Questions

The first step was to identify the relevant research questions for this SMS. The main Research Question (RQ) that guides this work is: **How do the tools reported in the literature use ML algorithms to assist in evaluating UX?** We established seven Sub-Research Questions (SQ) related to the identified tools to answer this RQ, in Table 1.

### 4.2 Databases

For the conduct of the SMS, we selected different databases for the collection of primary studies, including ACM Digital Library, Engineering Village, IEEE Xplore, Web of Science e Springer. We chose these bases because they provided an excellent search engine and

**Table 1: Sub-Research Questions.**

| # | Sub-question | Motivation |
|---|---|---|
| SQ01 | Which application platform is evaluated in the paper? | This question seeks to contextualize the tool within the current technological ecosystem. Knowing the target platform of the tool allows you to understand the design, development, and deployment constraints and opportunities. |
| SQ02 | What equipment is adopted for analysis? | Understanding which equipment is used for the study provides information about the complexity of data collection and the type of information the tool plans to extract. |
| SQ03 | What type of analysis was applied in the tool? | Defining the type of analysis applied in the tool helps to delimit the scope of the tool's assessment. |
| SQ04 | What type of data is used by the tool? | Understand what data the tools use to determine the capabilities, limitations, and if it suits a given context. |
| SQ05 | Which algorithm (classifier/ grouping) is most used? | Identify how the choice of algorithm directly impacts the way data is interpreted and the conclusions the tool can generate. |
| SQ06 | What metrics are used to evaluate the algorithm's learning? | This question aims to understand how the performance of the algorithm used by the tool is assessed, which impacts the interpretation of results and optimization. |
| SQ07 | What data visualization techniques are used by the tools? | Identify the techniques used by the tools allows to understand the information processed by them is made available to the user after processing. |

concentrated much of the studies that involve the topics of interest of this research, being recommended by different researchers in the conduct of secondary studies [20, 47, 55].

## 4.3 Search string

After the selection of databases, we performed a string search. We executed the search string in each of the digital libraries. As a parameter for this SMS, some primary studies were previously analyzed as control studies since they addressed the subject of SMS. These studies helped refine the string, which continued until appropriate studies were returned from previously selected study lists.

The string search used involved the different topics investigated in the study: user experience, machine learning, and tools. In this sense, variations of words that contemplated the topics resulted in the following string search:

("ux" OR "user experience" OR "user-centered evaluation" OR "user evaluation") **AND** ("machine learning" OR "sentiment analysis" OR "emotion detection") **AND** ("tool" OR "semi-automatic" OR "automatic" OR "software")

We searched by applying the metadata query from all sources to each study and adapting the search string syntax for implementation in each digital library. We only performed the search string in English.

## 4.4 Selection Criteria

The next step was to define the Inclusion Criteria (IC) and Exclusion Criteria (EC). We based the construction of the IC and EC on the guidelines of Kuhrmann et al. [34]. We included or excluded a study if it met any IC or EC positively, with the latter case resulting in its removal from the initial search. The inclusion and exclusion criteria can be seen in Table 2.

**Table 2: Selection Criteria.**

| ID | Description |
|---|---|
| **IC1** | Studies that present experimental studies on tools (methods or algorithms) that conduct UX evaluations automatically using ML techniques. |
| **IC2** | Studies that present tools (methods or algorithms) for sentiment/emotion analysis in interactive applications (desktop, mobile, or web). |
| **IC3** | Studies that present experimental studies on tools (methods or algorithms) for sentiment/emotion analysis in interactive applications (desktop, mobile, or web). |
| **EC1** | Duplicate studies (e.g., a study published in different venues or on different dates). Only the most complete and recent version will be considered in this case. |
| **EC2** | Studies that are not written in English. |
| **EC3** | The following types of studies: books, doctoral theses, master's theses, patents, tutorials, workshop proposals, or posters. |
| **EC4** | Studies that do not provide full-text access. |

## 4.5 Quality Criteria

We established Quality Criteria (QC) to assess the quality of the selected studies. The evaluation using QCs aims to quantify the relevance of each study. The QCs used in this study are described in Table 3.

**Table 3: Quality Criteria.**

| ID | Description |
|---|---|
| **QC1** | Does the study propose a tool that facilitates automated UX evaluation through the use of ML algorithms? |
| **QC2** | Does the tool support at least one option for modal analysis (voice, text, mouse, video, and other)? |
| **QC3** | Does the study report how the tool was applied in sentiment analysis? |
| **QC4** | Is the use of the tool clearly/detailed described? |

Each of the studies was evaluated by each researcher, with the following scores assigned to the QCs: **Yes (Y): 1.0; Partially (P): 0.5; No (N): 0.0**. Thus, each primary study had a final score that could range from 0.0 (minimum score) to 4.0 (maximum score). In this sense, score intervals were defined to validate the studies' quality. If the study had a final score less than or equal to 0.5, it was considered **"Weak"**; between 0.6 and 1.5, it was considered **"Fair"**; between 1.6 and 2.5, it was considered **"Good"**; between 2.6 and 3.5, it was considered **"Very Good"**; and above 3.5, it was considered **"Excellent"**, as suggested by [31]. In the context of this SMS, studies with a quality score lower than 1.5 were excluded from the final list of selected primary studies, even if they belonged to the research domain.

## 4.6 Data Extraction

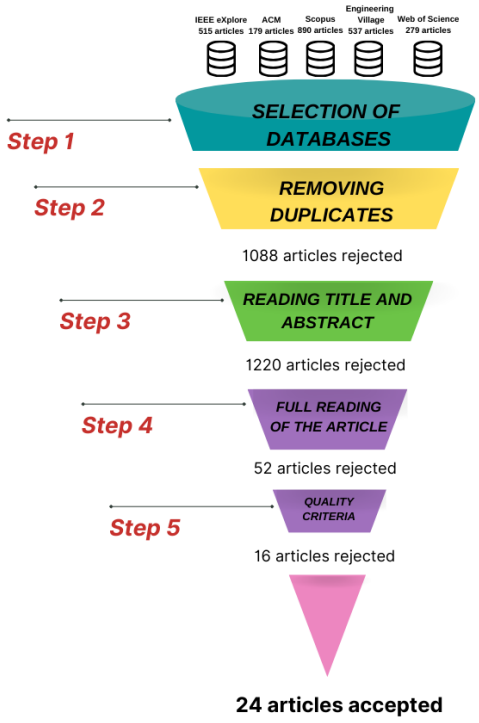
After evaluating the quality criteria, we created a spreadsheet in Google Sheets containing only the information necessary to answer the SQs to organize the selected studies. We extracted the following data: evaluated platform, equipment used, applied analysis, type of data used by the tool, adopted algorithm, metrics used, and data visualization techniques employed. We collected all this information from a list with pre-selected options.

## 4.7 Database results

Figure 2 summarizes the entire study selection process. The search returned 2400 studies. All potential studies were accessed using the internet connection and authentication provided by the first

author's university. When a potential article was not available free of charge on the institution's network, we contacted the corresponding author to request a preprint. Some studies were available in different digital libraries, but we considered them only once, resulting in 1.312 studies that were considered valid. We adopted the *Parsif.al* platform as a support tool in selecting studies.

**Figure 2: Results of the SMS filters.**




This SMS involved four researchers to reduce the interpretation bias of a single researcher. Three doctoral researchers reviewed the protocol and the inclusion and exclusion criteria and analyzed the research strategy. To evaluate the level of reliability of the study selection process, two researchers independently classified a sample of 60 studies randomly selected from the set of returned to measure the level of agreement between them. In this classification, we evaluated the title and summary of each study and classified it based on the selection criteria. After this step, we applied the statistical test of Fleiss et al. [24], a measure of intra and inter-observer agreement that evaluates the degree of agreement beyond what would be expected only by chance. The evaluation result showed a Kappa agreement = 0.89 among the researchers, representing an almost perfect agreement. From this result, we continued the steps of selection of studies.

Using the title and abstract (first filter) as a basis, we realized that 1.220 studies not aligned with the research objective were rejected, resulting in 92 selected studies. In the next step (second filter), we did a complete reading of the studies included in the first filter, applying the same inclusion and exclusion criteria. This step aimed to ensure a more accurate analysis of the studies. The results were analyzed and discussed, and we resolved discrepancies among the researchers. After this penultimate filter, 40 studies remained, which were then submitted to the quality criteria. In the final step, applying quality criteria, only those studies that could contribute to the study theme were selected, eliminating 16 studies that did not meet the criteria. At the end of the process, we accepted 24 studies.

## 5 RESULTS

Table 4 presents the 24 selected studies, which include the study ID (first column), authors with study reference (second column), year of publication (third column), type of publication (journal or conference), and the result of QCs.

**Table 4: Studies selected based on QCs.**

| ID | References | Year | Type | Quality evaluation | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | QC1 | QC2 | QC3 | QC4 | Value | Status |
| S01 | Sanchis-Font et al. [62] | 2021 | Journal | S | P | P | P | 2,5 | Accepted |
| S02 | Sánchez et al. [61] | 2015 | Journal | S | S | P | P | 3,0 | Accepted |
| S03 | Piskioulis et al. [56] | 2021 | Conference | P | S | P | P | 2,5 | Accepted |
| S04 | Hedegaard and Simonsen [28] | 2013 | Conference | S | S | P | P | 3,0 | Accepted |
| S05 | Liapis et al. [38] | 2019 | Conference | S | P | P | P | 2,5 | Accepted |
| S06 | Fan et al. [22] | 2020 | Conference | P | P | P | P | 2,0 | Accepted |
| S07 | Jabbar et al. [32] | 2019 | Conference | P | P | P | P | 2,0 | Accepted |
| S08 | Song et al. [64] | 2021 | Conference | P | S | P | N | 2,0 | Accepted |
| S09 | Bakar and Seong [5] | 2020 | Journal | P | S | P | P | 2,5 | Accepted |
| S10 | Ajenaghughrure et al. [2] | 2019 | Conference | N | S | P | P | 2,0 | Accepted |
| S11 | Franco et al. [25] | 2020 | Conference | S | S | P | S | 3,5 | Accepted |
| S12 | Zhang et al. [69] | 2019 | Journal | N | P | S | S | 2,5 | Accepted |
| S13 | Maalej et al. [42] | 2016 | Journal | S | S | N | P | 2,5 | Accepted |
| S14 | Alkalbani et al. [3] | 2016 | Conference | S | P | N | N | 1,5 | Accepted |
| S15 | Desolda et al. [16] | 2021 | Conference | S | P | P | N | 2,0 | Accepted |
| S16 | Munim et al. [51] | 2017 | Conference | N | P | S | S | 2,5 | Accepted |
| S17 | Soure et al. [65] | 2021 | Journal | S | S | P | P | 3,0 | Accepted |
| S18 | Bakiu and Guzman [6] | 2021 | Journal | P | P | P | P | 2,0 | Accepted |
| S19 | Liapis et al. [37] | 2021 | Journal | N | P | P | S | 2,0 | Accepted |
| S20 | da Silva Franco et al. [15] | 2019 | Journal | S | S | S | S | 4,0 | Accepted |
| S21 | Fan et al. [23] | 2021 | Journal | S | P | N | P | 2,0 | Accepted |
| S22 | Batch et al. [7] | 2024 | Journal | S | S | S | S | 4,0 | Accepted |
| S23 | Drungilas et al. [19] | 2024 | Journal | S | P | S | S | 3,5 | Accepted |
| S24 | Liang et al. [36] | 2023 | Conference | S | P | S | S | 3,5 | Accepted |
| S25 | Kaiser and Stern | 2021 | Conference | N | N | P | N | 0,5 | Rejected |
| S26 | Saura et al. | 2018 | Journal | N | N | N | N | 0,0 | Rejected |
| S27 | Gomes et al. | 2018 | Conference | N | N | N | N | 0,0 | Rejected |
| S28 | Bernardo et al. | 2015 | Journal | N | N | N | N | 0,0 | Rejected |
| S29 | Bernardo et al. | 2017 | Conference | N | P | N | N | 0,5 | Rejected |
| S30 | Fransisca et al. | 2021 | Conference | N | P | N | N | 0,5 | Rejected |
| S31 | Yu et al. | 2019 | Conference | N | N | P | N | 0,5 | Rejected |
| S32 | Jansen and Colombo | 2022 | Conference | N | N | N | N | 0,0 | Rejected |
| S33 | Biringa and Kul | 2021 | Conference | N | N | N | N | 0,0 | Rejected |
| S34 | Liapis et al. | 2019 | Conference | N | N | P | N | 0,5 | Rejected |
| S35 | Liapis et al. | 2021 | Conference | N | P | N | N | 0,5 | Rejected |
| S36 | Garcia-Garcia et al | 2018 | Journal | N | N | N | N | 0,0 | Rejected |
| S37 | Cong et al. | 2022 | Journal | N | N | N | N | 0,0 | Rejected |
| S38 | Lim et al. | 2014 | Conference | N | N | N | N | 0,0 | Rejected |
| S39 | Orozco-Mora et al. | 2022 | Journal | N | P | N | N | 0,5 | Rejected |
| S40 | Balamurali et al. | 2023 | Conference | N | N | P | P | 1,0 | Rejected |

Table 5 summarizes the analyses performed for the specific SQs of this SMS.

### 5.1 SQ1 - Supported platform

The analysis of primary studies revealed more frequent support for Web and Mobile applications, as evidenced in Table 5, which represents 20 studies that mention or develop tools used to evaluate. For example, S22 [7] proposed a client/server system that extracts user behavior from videos, incorporating a computational backend and an interactive interface on the web. This system allows qualitative and quantitative user experience analysis, using time visualization techniques to highlight interest segments. In addition, it adopts an asynchronous workflow, allowing background processing and the transmission of results back to the customer as they become available.

**Table 5: SMS results for each sub-question.**

| SQs | Answers | # | Studies |
|---|---|---|---|
| SQ1 | Mobile | 11 | S01-S03, S07, S10, S12-S14, S17-S19 |
| | Web | 9 | S06, S09, S11, S14, S15, S17, S19, S20, S23 |
| | Desktop | 6 | S02, S14, S16, S17, S19, S20, S23 |
| | Does not mention | 5 | S04, S05, S08, S21, S24 |
| SQ2 | Computer (to capture inputs) | 9 | S09, S10, S14, S16-S18, S20, S21, S23 |
| | Camera | 7 | S11, S15-S17, S20, S22, S23 |
| | Does not specify | 7 | S01, S04, S07, S11, S12, S13, S24 |
| | Microphone | 6 | S06, S08, S11, S17, S20, S22 |
| | Equipments of physiological data | 3 | S05, S10, S19 |
| | Smartphone | 2 | S02, S03 |
| | Eye Tracker | 2 | S11, S20 |
| | Computer (for data processing) | 2 | S06, S22 |
| | Laboratory equipped | 1 | S05 |
| | Web system | 1 | S22 |
| | Kinect | 1 | S02 |
| | Mouse Tracker | 1 | S20 |
| SQ3 | Sentiment analysis | 14 | S01-S03, S05, S07-S09, S13, S14, S18, S21-S24 |
| | User eXperience | 10 | S01, S02, S05, S08, S10, S18, S19, S21-S23 |
| | Usability | 7 | S02, S05, S06, S18, S21-S23 |
| | Sentiments and UX | 5 | S11, S12, S15, S16, S20 |
| | Usability and UX | 1 | S04 |
| SQ4 | Plain text | 12 | S01, S04, S06, S07, S11, S13, S14, S17, S18, S21, S22, S24 |
| | Audio e video | 4 | S11, S17, S20, S22 |
| | Audio | 3 | S06, S08, S21 |
| | Physiological data | 3 | S05, S10, S19 |
| | Eye tracker | 2 | S11, S20 |
| | Screenshot | 2 | S12, S15 |
| | Gyroscope e accelerometer | 1 | S03 |
| | Keyboard inputs | 1 | S09 |
| | Video | 4 | S02, S12, S16, S23 |
| | Mouse Tracking | 1 | S20 |
| | Audio | 3 | S06, S08, S21 |
| | Screen recording | 1 | S16 |
| SQ5 | SVM | 9 | S04, S06, S07, S10, S11, S14, S18, S20, S21 |
| | CNN | 8 | S01, S06, S08, S11, S20-S23 |
| | Does not mention | 5 | S02, S09, S12, S16, S17 |
| | Random forest | 4 | S06, S10, S15, S21 |
| | Decision tree | 3 | S03, S13, S15 |
| | RNN | 3 | S06, S21, S22 |
| | Logistic regression | 2 | S03, S10 |
| | SVM-Adapted | 2 | S05, S19 |
| SQ6 | Recall | 11 | S01, S04, S06-S08, S13, S14, S18, S19, S21, S24 |
| | Accuracy | 10 | S01, S03, S05, S10, S14, S15, S19, S20, S23, S24 |
| | Precision | 10 | S01, S04, S06, S07, S13, S14, S18, S19, S21, S24 |
| | F1 | 9 | S01, S04, S06, S07, S13, S18, S19, S21, S24 |
| | Does not mention | 6 | S02, S09, S11, S12, S17, S22 |
| | AUC | 2 | S07, S19 |
| | ROC curve | 1 | S10 |
| | Efficiency | 1 | S16 |
| | Efficacy | 1 | S16 |
| | True Positive Rate | 1 | S05 |
| | False positive rates | 1 | S05 |
| | Kappa | 1 | S19 |
| SQ7 | Graphics/Dashboard | 19 | S02, S03, S06, S07, S09-S14, S16-S24 |
| | Tabel | 5 | S04-S06, S10, S12 |
| | Word cloud | 2 | S11, S20 |
| | Scanpath | 2 | S11, S20 |
| | Heatmap | 2 | S15, S23 |
| | Mel-Spectogram | 1 | S08 |
| | Emojitext | 1 | S20 |
| | Does not mention | 1 | S01 |

The study S02 [61] exemplified another mobile platform. This study describes an extensible software platform called Vikara, which facilitates the integration of various emotion analysis tools into applications. Vikara offers uniform interfaces and services for applications to access the results of the tools, including a FACS-based emotion analysis Facial Action Coding System using Kinect and a self-report interface for Android mobile devices. Regarding desktop platforms, we can cite the study S20 [15], which presents the tool Uxmood. The tool allows for joint analysis of usability and user experience, using video data, audio, interaction logs, eye trackers, and sentiment analysis techniques applied to video content, audio, and transcribed text to obtain insights about participants' user experience.

## 5.2 SQ2 - Data collection equipment

Several categories of equipment necessary for data collection by tools were identified, as presented in Table 5. The 'Computer' category covers data collected from system inputs, such as mouse and keyboard, as well as screenshots and text taken from the internet. In addition, different types of collection are highlighted, such as devices for physiological data. For example, study S05 [38] explores stress detection through physiological signals such as electrodermal activity and skin temperature. This data is essential to understanding users' emotional responses during interaction with digital systems.

Cameras are also often used to record users' facial expressions, capturing their emotional reactions during interaction with the system. For example, the S02 study [61] describes an emotion-sensing platform that uses cameras to capture users' facial expressions and other data, such as voice and posture, to infer the emotional state during interaction with a system. Another relevant category is the eye tracker, present in the study S20 [15], which collects visual data about user eye movement. The captured data, such as x and y coordinates and timestamps, are stored in files in .csv, filtered to remove errors outside the visible field of the monitor and ensure correct synchronization with the application under test.

## 5.3 SQ3 - Types of analysis applied

Regarding the type of analysis the tools performed, we observed that researchers more often utilized these tools for sentiment analysis and UX evaluation, enabling the discovery of studies at the intersection of both areas, as in study S09[5]. In the category 'Sentiment Analysis,' study S01 [62] examines cross-domain polarity models to evaluate UX in virtual learning environments, using ML to analyze user opinions on learning platforms. Already in the category 'Usability and UX, ' the study S04 [27] investigates the extraction of usability and UX information from user reviews of online software and video games. The study S16 [51] falls into the category 'Sentiments and UX'; the focus is on developing a UX assessment tool that uses emotion detection through facial expressions. In particular, in the S20 study [14], the authors present a tool that combines video data, audio, logs of interaction, and eye trackers to analyze the UX comprehensively. It uses audio and speech transcriptions to get insights about the UX.

## 5.4 SQ4 - Types of data used

We identified 12 categories of data for the data used by the tools. A category that predominated was pure text, with 12 studies adopting this format for easy collection and analysis. The study S06 [22] exemplifies this practice. The authors used Natural Language Processing (NLP) algorithms to analyze feedback text in think-aloud sessions. The treatment of the pure text began with the manual transcription of the sessions and the segmentation into smaller parts

based on pauses and the meaning of verbalizations, facilitating annotation and analysis. According to the authors, each segment was categorized into Reading, Procedure, Observation, and Explanation classes and analyzed for negations, questions, and feelings. Pre-processing includes the removal of stopwords, tokenization, and TF-IDF calculations and embedding vectors of words. After feature extraction, ML tools are trained with these feature vectors and their labels.

In addition, the analysis of audio and video data is present. The S06 study [22], for example, uses think-aloud audios to train an ML model capable of automatically detecting usability issues by segmenting audio into 10-second sections and extracting acoustic features like pitch, energy, formats, and Mel-frequency cepstral coefficients (MFCCs), which are processed to feed the model. The study S17 [65] used video data to analyze user interaction with a think-aloud analysis tool. The video records users' actions in the interface, such as clicks, scrolls, and typing, and serves as a basis for identifying behavior patterns and usability issues. Additionally, the video allows users to observe their body language, such as facial expressions, gestures, and posture, providing insights into their emotions and level of engagement with the tool.

## 5.5 SQ5 - Algorithms adopted

The most common algorithm authors adopted was SVM (Support Vector Machine), as demonstrated in the S06 study [22]. The authors use SVM to automatically detect usability problems in think-aloud sessions, combining textual and acoustic characteristics to train the model. The S19 study [37] explores the use of SVM to detect stress from physiological signals, such as skin conductance. The SVM stands out for its effectiveness in classifying text and numerical data, especially on moderate-sized datasets.

In addition to SVM, other ML algorithms are mentioned in the studies. CNN (Convolutional Neural Network), a popular algorithm for image processing and sequential data analysis, is cited in six studies. In S20 [14], CNN classifies users' facial expressions in videos, inferring their emotions during system interaction. Study S21 [23] highlights CNN as one of the algorithms used to detect usability problems from think-aloud data.

Other algorithms, such as Decision Trees, are employed in the S03 [56] study to detect emotions from accelerometer and keyboard data, while Random Forest, present in the S06 [22], S10 [2], S15 [16] and S21 [23], stands out for its robustness in data classification. For example, the S06 [22] study compared the performance of Random Forest with the SVM in detecting usability issues, while the S10 study [2] constructed a model to predict the level of trust of users in technologies.

These algorithms are chosen because they can handle different types of data and are effective in identifying complex patterns, which is critical for accurately assessing UX. The literature's predominance of SVM and CNN shows their relevance and adaptability in UX evaluation tools that use ML.

## 5.6 SQ6 - Metrics adopted by tools

The most commonly used metrics in evaluating the performance of ML tool algorithms are F1, accuracy, precision, and recall. For example, study S21 [23] employs the metrics of accuracy, precision, and recall to evaluate its model.

It is worth mentioning that accuracy is a widely used metric that indicates the proportion of correct ratings about the total number of ratings achieved [63]. However, accuracy can be misleading in unbalanced datasets because models can reach high values simply by predicting most instances as the majority class. These results highlight the importance of complementary metrics like precision and recall. Accuracy is calculated by the number of correctly classified positive examples divided by the number of examples labeled by the system as positive [63]. It indicates the proportion of correct positive predictions, which is crucial to evaluating the quality of the model's positive predictions. Recall measures the proportion of true positives identified in the total number of positive cases [63]. This metric is essential when identifying all positive cases, reflecting the model's ability to capture all relevant instances. F1 is the harmonic mean between precision and recall, offering a balance between the two metrics [63]. It is beneficial when seeking balanced performance in both measures, providing a single metric that synthesizes the model's effectiveness in accuracy and coverage of predictions.

## 5.7 SQ7 - Data visualization techniques

Among the visualization techniques employed, it is observed that most of the studies used graphics and/or dashboards, as demonstrated in the study S22 [7]. The dashboard proposed in S22 combines different visualization techniques to present an integrated view of the data, including line charts to represent the timeline of emotions, stacked bar charts to show the duration of each feeling, word clouds to highlight the most common terms, and scan paths to visualize the user's eye movement in the interface. In addition to these, other forms of representation are identified, such as tables, heatmaps, and word clouds. However, the sum of these categories does not equal or exceed the category of graphs and/ or dashboards, showing the predominance of this visualization technique in the literature analyzed.

# 6 DISCUSSION

The results of this study contribute to the identification of tools that use ML techniques to support the evaluation of UX. These tools demonstrate the potential to assist professionals, reduce their workload, identify known patterns, and optimize their operation time.

The analysis of support platforms for UX evaluation applications revealed a predominance of mobile systems, allowing data collection in real scenarios and taking advantage of the ubiquitous nature of smartphones. This preference can be attributed to the ability of mobile devices to integrate sensors that other platforms do not have by nature, such as accelerometers and cameras, which collect behavioral and emotional data during user interactions. While web platforms offer flexibility and accessibility for remote analytics and collaboration, UX evaluation through mobile platforms allows for rich contextual data collection, capturing the user experience in real-world situations. The use of web and desktop platforms remains relevant for tasks that require more complex data processing

or large-scale collaboration. However, the rise of mobile platforms highlights their growing importance in UX evaluation.

Data collection for UX assessment benefits from various equipment, ranging from system input devices such as mouse and keyboard to more sophisticated tools like eye trackers and physiological data capture devices. This diversity reflects the search for a global understanding of user experience, incorporating interactions with the interface, emotional reactions, and cognitive state during system use. Different equipment allows the collection of complementary data, enriching the analysis and providing more in-depth insights about UX.

UX analysis tools employ a variety of techniques, with an emphasis on the analysis of feelings and usability assessment. Based on NPL and ML, analyzing feelings allows identifying the polarity and intensity of emotions users express in their feedback text. The usability evaluation focuses on identifying interaction problems using usability tests, think-aloud, and heuristic evaluation. As observed in some studies, combining these techniques allows a more complete assessment of UX, considering both the emotional and practical aspects of interaction with the system.

The diversity of data types used by UX analysis tools highlights the importance of integrating different sources of information to build a comprehensive assessment. For its ease of collection and analysis, pure text is widely used to extract perceptions about UX from online reviews, comments on social networks, and transcriptions of think-aloud sessions. Audio and video data, in turn, allow the analysis of nuances of verbal and non-verbal communication, capturing information on intonation, facial expressions, gestures, and posture. The integration of data from eye tracking enables the analysis of the visual attention pattern of users, identifying areas of interest, navigation difficulties, and elements that generate confusion or frustration.

Applying ML algorithms such as SVM, CNN, and Random Forest automates the analysis of UX data, streamlining the process and revealing complex patterns that would be difficult to identify manually. The choice of the most suitable algorithm depends on the characteristics of the data and the research objectives. It is important to highlight that the use of ML algorithms does not replace the expertise of the UX evaluator but serves as an auxiliary tool, providing insights that can guide analysis and decision-making.

The performance evaluation of ML algorithms is based on metrics such as accuracy, precision, recall, and F1, which provide quantitative indicators of prediction quality. The most relevant metrics depend on the problem and research objectives. Accuracy, for example, can be a misleading metric in unbalanced data sets. At the same time, precision and recall provide more specific information about the model's ability to correctly identify positive and negative cases.

Data visualization techniques, such as graphics, tables, and dashboards, are key in communicating the results of UX evaluation, making data more understandable and accessible to different audiences. Interactive dashboards, which combine various types of views and offer data filtering and exploration capabilities, make it easy to identify patterns, trends, and insights relevant to UX optimization. The choice of the most appropriate visualization technique depends on the data type, the analysis's complexity, and the target audience.

## 7 CONCLUSION

This study sought to understand which ML tools are available to support UX analysis. To identify these tools, we conducted an SMS that allowed us to classify and categorize the tools found. We expected that the results would guide researchers and professionals in selecting appropriate tools for their specific contexts and identify gaps and opportunities for future research.

The detailed analysis of the studies revealed significant trends in developing and implementing these tools. They are predominantly designed for mobile platforms, although there is considerable support for web and desktop applications. The equipment used for data collection was significantly diverse, ranging from conventional computers for capturing fundamental interactions to specialized devices such as eye trackers and physiological sensors. This variety of instruments allows for a more comprehensive and multifaceted user experience analysis, incorporating objective and subjective data into the assessment.

In terms of analysis, the tools focus primarily on sentiment analysis and user experience assessment, often combining both approaches for a more holistic understanding. The data used are predominantly textual, although there is considerable integration of multimodal data, including audio, video, and physiological metrics. This diversity of data sources allows for a more robust and contextualized user experience analysis.

As for the technical aspects, Support Vector Machine (SVM) emerges as the most used algorithm, followed by Convolutional Neural Networks (CNN) and Random Forest, demonstrating a preference for established and reliable machine learning methods. The evaluation of these algorithms is mainly performed through traditional metrics such as F1, accuracy, precision, and recall, providing a solid basis for the validation of results. Data visualization is predominantly performed through graphs and dashboards, facilitating the interpretation and analysis of results by researchers and professionals in the field.

The results show that UX assessment constantly evolves, with a clear trend towards more integrated and holistic approaches. Combining different data sources, analysis methods, and visualization techniques indicates a mature field with ample potential for innovation, especially in integrating new technologies and analysis methods. The predominance of mobile platforms highlights the growing importance of this technology in data collection in real scenarios, taking advantage of smartphones' ubiquity and advanced sensors.

Additionally, the diversity of equipment used, ranging from traditional input devices such as mouse and keyboard to sophisticated tools such as eye trackers and physiological data capture devices, reflects the quest for a comprehensive understanding of user experience. Integrating behavioral, emotional, and cognitive data enriches analysis and provides deeper insights into UX. The analysis techniques used, such as sentiment analysis and usability evaluation, combined with ML algorithms like SVM, CNN, and Random Forest, demonstrate the effectiveness of automation in identifying complex patterns and optimizing the evaluation process. The metrics used to evaluate algorithms, including accuracy, precision, recall, and F1-Score, provide essential quantitative indicators to measure the quality of predictions and the effectiveness of applied models. The

results of this study can serve as a valuable guide for researchers and professionals who seek to implement or develop UX analysis tools based on machine learning.

This panorama suggests that our work is aligned with the Great Challenges of Research in Information Systems in Brazil for the decade 2016-2026 [4], especially in Challenge 4, which addresses the Socio-technical Perspective of Information Systems. This challenge emphasizes that information systems integrate people and technology entirely, with multiple relationships emerging from this interaction. The UX analysis tools based on ML identified in this study exemplify perfectly this socio-technical vision, seeking to understand and improve the complex interaction between users and systems through interdisciplinary methods that combine technical aspects (ML, sensors, data processing) with human elements (experience, emotions, behavior).

The results of this study provide a comprehensive view of current tools for using ML for UX evaluation and pave the way for future innovations that will transform how we understand and improve the user experience in information systems.

## ACKNOWLEDGMENTS

The authors thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001 and the Universidade Federal do Pampa (UNIPAMPA - Alegrete) for their support. Williamson Silva thanks FAPERGS for its financial support (Project ARD/ARC – process 22/2551-0000606-0). Pedro Henrique Dias Valle thanks FAPEMIG for its financial support (Process APQ-00743-22). Gabriel Machado Lunardi thanks for CNPq 402086/2023-6 and FAPERGS ARD/ARC – process 24/2551-0000645-1.

## REFERENCES

[1] Abdallah M. H. Abbas, Khairil Imran Ghauth, and Choo-Yee Ting. 2022. User Experience Design Using Machine Learning: A Systematic Review. *IEEE Access* 10 (2022), 51501–51514. https://doi.org/10.1109/ACCESS.2022.3173289

[2] Ighoyota Ben Ajenaghughrure, Sonia C Sousa, Ilkka Johannes Kosunen, and David Lamas. 2019. Predictive model to assess user trust: a psycho-physiological approach. In *Proceedings of the 10th Indian conference on human-computer interaction*. 1–10.

[3] Asma Musabah Alkalbani, Ahmed Mohamed Ghamry, Farookh Khadeer Hussain, and Omar Khadeer Hussain. 2016. Sentiment analysis and classification for software as a service reviews. In *2016 IEEE 30th international conference on advanced information networking and applications (AINA)*. IEEE, 53–58.

[4] Renata Araújo and R Suzana. 2017. Grand research challenges in information systems in brazil 2016–2026. *Brazilian Computer Society. Clodis Boscarioli Renata Araujo and Rita Suzana* (2017), 2016–2026.

[5] Zuriana Abu Bakar and Ong Sze Seong. 2020. The Development of Web-Based Emotion Detection System Using Keyboard Actions (EDS-KA). *International Journal on Advanced Science, Engineering and Information Technology* (2020). https://api.semanticscholar.org/CorpusID:216403850

[6] Elsa Bakiu and Emitza Guzman. 2017. Which feature is unusable? Detecting usability and user experience issues from user reviews. In *2017 IEEE 25th international requirements engineering conference workshops (REW)*. IEEE, 182–187.

[7] A. Batch, Y. Ji, M. Fan, J. Zhao, and N. Elmqvist. 2024. uxSense: Supporting User Experience Analysis with Visualization and Computer Vision. *IEEE Transactions on Visualization and Computer Graphics* 30, 07 (jul 2024), 3841–3856. https://doi.org/10.1109/TVCG.2023.3241581

[8] Regina Bernhaupt, Philippe Palanque, Dimitri Drouet, and Celia Martinie. 2019. Enriching task models with usability and user experience evaluation data. In *Human-Centered Software Engineering: 7th IFIP WG 13.2 International Working Conference, HCSE 2018, Sophia Antipolis, France, September 3–5, 2018, Revised Selected Papers 7*. Springer, 146–163.

[9] Kim Carmona, Erin Finley, and Meng Li. 2018. The Relationship Between User Experience and Machine Learning. *Available at SSRN 3173932* (2018).

[10] Helena de Medeiros Caseli and Maria das Graças Volpe Nunes. 2023. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*.

[11] Michael Chromik, Florian Lachner, and Andreas Butz. 2020. ML for UX? - An Inventory and Predictions on the Use of Machine Learning Techniques for UX Research. 1–11. https://doi.org/10.1145/3419249.3420163

[12] Michael Chromik, Florian Lachner, and Andreas Butz. 2020. Ml for ux?-an inventory and predictions on the use of machine learning techniques for UX research. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. 1–11.

[13] Ricardo Luiz Hentges Costa, Tales Schifelbein Soares, Gabriel Machado Lunardi, Pedro Henrique Dias Valle, and Williamson Silva. 2024. Professionals' Perceptions of the Interaction between User Experience and Machine Learning. In *Proceedings of the 20th Brazilian Symposium on Information Systems*. 1–9.

[14] Roberto Yuri da Silva Franco, Rodrigo Santos do Amor Divino Lima, Rafael do Monte Paixão, Carlos Gustavo Resque dos Santos, and Bianchi Serique Meiguins. 2019. Uxmood—a sentiment analysis and information visualization tool to support the evaluation of usability and user experience. *Information* 10, 12 (2019), 366.

[15] Roberto Yuri da Silva Franco, Rodrigo Santos do Amor Divino Lima, Rafael do Monte Paixão, Carlos Gustavo Resque dos Santos, and Bianchi Serique Meiguins. 2019. UXmood—A Sentiment Analysis and Information Visualization Tool to Support the Evaluation of Usability and User Experience. *Information* 10, 12 (2019). https://doi.org/10.3390/info10120366

[16] Giuseppe Desolda, Andrea Esposito, Rosa Lanzilotti, and Maria F Costabile. 2021. Detecting emotions through machine learning for automatic UX evaluation. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18*. Springer, 270–279.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

[18] ISO DIS. 2009. 9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems. *International Standardization Organization (ISO). Switzerland* (2009).

[19] Darius Drungilas, Ignas Ramašauskas, and Mindaugas Kurmis. 2024. Emotion Recognition in Usability Testing: A Framework for Improving Web Application UI Design. *Applied Sciences* 14, 11 (2024). https://doi.org/10.3390/app14114773

[20] Tore Dyba, Torgeir Dingsoyr, and Geir K Hanssen. 2007. Applying systematic reviews to diverse study types: An experience report. In *First international symposium on empirical software engineering and measurement (ESEM 2007)*. IEEE, 225–234.

[21] Katti Faceli, Ana Carolina Lorena, João Gama, Tiago Agostinho de Almeida, and Andre Carlos Ponce de Leon Ferreira de Carvalho. 2021. *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC.

[22] Mingming Fan, Yue Li, and Khai N Truong. 2020. Automatic detection of usability problem encounters in think-aloud sessions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 2 (2020), 1–24.

[23] Mingming Fan, Ke Wu, Jian Zhao, Yue Li, Winter Wei, and Khai N. Truong. 2020. VisTA: Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 343–352. https://doi.org/10.1109/TVCG.2019.2934797

[24] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. John Wiley & Sons.

[25] Roberto Yuri Da Silva Franco, Alexandre Abreu De Freitas, Rodrigo Santos Do Amor Divino Lima, Marcelle Pereira Mota, Carlos Gustavo Resque Dos Santos, and Bianchi Serique Meiguins. 2019. UXmood-A tool to investigate the user experience (UX) based on multimodal Sentiment analysis and information visualization (InfoVis). In *2019 23rd International Conference Information Visualisation (IV)*. IEEE, 175–180.

[26] Marc Hassenzahl. 2018. The thing and I: understanding the relationship between user and product. *Funology 2: from usability to enjoyment* (2018), 301–313.

[27] Steffen Hedegaard and Jakob Grue Simonsen. 2013. Extracting usability and user experience information from online user reviews. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2089–2098. https://doi.org/10.1145/2470654.2481286

[28] Steffen Hedegaard and Jakob Grue Simonsen. 2013. Extracting usability and user experience information from online user reviews. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2089–2098.

[29] Kasper Hornbæk and Morten Hertzum. 2017. Technology Acceptance and User Experience: A Review of the Experiential Component in HCI. *ACM Trans. Comput.-Hum. Interact.* 24, 5, Article 33 (oct 2017), 30 pages. https://doi.org/10.1145/3127358

[30] Jamil Hussain, Wajahat Ali Khan, Taeho Hur, Hafiz Syed Muhammad Bilal, Jaehun Bang, Anees Ul Hassan, Muhammad Afzal, and Sungyoung Lee. 2018. A multimodal deep log-based user experience (UX) platform for UX evaluation. *Sensors* 18, 5 (2018), 1622.

[31] Aníbal Iung, João Carbonell, Luciano Marchezan, Elder Rodrigues, Maicon Bernardino, Fabio Paulo Basso, and Bruno Medeiros. 2020. Systematic mapping study on domain-specific language development tools. *Empirical Software*

*Engineering* 25 (2020), 4205–4249.
[32] Jahanzeb Jabbar, Iqra Urooj, Wu JunSheng, and Naqash Azeem. 2019. Real-time sentiment analysis on E-commerce application. In *2019 IEEE 16th international conference on networking, sensing and control (ICNSC)*. IEEE, 391–396.
[33] Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. (2007).
[34] Marco Kuhrmann, Daniel Méndez Fernández, and Maya Daneva. 2017. On the pragmatic design of literature studies in software engineering: an experience-based guideline. *Empirical software engineering* 22, 6 (2017), 2852–2891.
[35] Karl Kurbel and B. Ulrich. 2008. *The making of information systems: Software engineering and management in a globalized world*. Springer. 1–591 pages. https://doi.org/10.1007/978-3-540-79261-1
[36] Feng Liang, Fang Hou, Siamak Farshidi, Slinger Jansen, et al. 2023. Sentiment analysis for software quality assessment. In *CEUR Workshop Proceedings*, Vol. 3567. CEUR WS, 17–24.
[37] Alexandros Liapis, Evanthia Faliagka, Christos P. Antonopoulos, Georgios Keramidas, and Nikolaos Voros. 2021. Advancing Stress Detection Methodology with Deep Learning Techniques Targeting UX Evaluation in AAL Scenarios: Applying Embeddings for Categorical Variables. *Electronics* 10, 13 (2021). https://doi.org/10.3390/electronics10131550
[38] Alexandros Liapis, Christos Katsanos, Nikos Karousos, Michalis Xenos, and Theofanis Orphanoudakis. 2019. UDSP+ stress detection based on user-reported emotional ratings and wearable skin conductance sensor. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 125–128.
[39] Cesar Eduardo Abud Limas. 2009. Sistemas integrados de gestão-ERP: benefícios esperados e problemas encontrados na implantação em pequenas empresas brasileiras. (2009).
[40] Danilo Lira, Gabriel Mac'Hamilton, Gabriel Stadtler, Myllena Almeida, Simone C dos Santos, and Jéssyka Vilela. 2023. The Use of Users' Personal Data to Improve Online Retail Services: An Analysis from a Systematic Literature Review. In *Proceedings of the XIX Brazilian Symposium on Information Systems*. 78–85.
[41] Yuwen Lu, Yuewen Yang, Qinyi Zhao, Chengzhi Zhang, and Toby Jia-Jun Li. 2024. AI Assistance for UX: A Literature Review Through Human-Centered AI. *arXiv preprint arXiv:2402.06089* (2024).
[42] Walid Maalej, Zijad Kurtanović, Hadeer Nabil, and Christoph Stanik. 2016. On the automatic classification of app reviews. *Requirements Engineering* 21 (2016), 311–331.
[43] Maylon Macedo. 2024. UX Data Visualization: Supporting Software Professionals in Exploring. In *Engineering Interactive Computer Systems: EICS 2023 International Workshops and Doctoral Consortium: Swansea, UK, June 26-27, 2023, Selected Papers*, Vol. 14517. Springer Nature, 207.
[44] C. Manning and H. Schutze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press. https://books.google.com.br/books?id=YiFDxbEX3SUC
[45] Christopher D Manning. 2009. *An introduction to information retrieval*.
[46] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal* 5, 4 (2014), 1093–1113.
[47] Emilia Mendes, Claes Wohlin, Katia Felizardo, and Marcos Kalinowski. 2020. When to update systematic literature reviews in software engineering. *Journal of Systems and Software* 167 (2020), 110607.
[48] M Moellendorff, Marc Hassenzahl, and Axel Platz. 2006. Dynamics of user experience: How the perceived quality of mobile phones changes over time. In User experience—towards a unified view. 74–78.
[49] Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*. Elsevier, 201–237.
[50] Ruth Mugge, Hendrik NJ Schifferstein, and Jan PL Schoormans. 2005. Product attachment and product lifetime: The role of personality congruity and fashion. *ACR European Advances* (2005).
[51] Kazi Md Munim, Iyolita Islam, Mahmuda Khatun, Md Mahboob Karim, and Muhammad Nazrul Islam. 2017. Towards developing a tool for UX evaluation using facial expression. In *2017 3rd international conference on Electrical Information and Communication Technology (EICT)*. IEEE, 1–6.
[52] Mie Nørgaard and Kasper Hornbæk. 2006. What do usability evaluators do in practice? an explorative study of think-aloud testing. In *Proceedings of the 6th Conference on Designing Interactive Systems* (University Park, PA, USA) *(DIS '06)*. Association for Computing Machinery, New York, NY, USA, 209–218. https://doi.org/10.1145/1142405.1142439
[53] Martin Obaidi, Lukas Nagel, Alexander Specht, and Jil Klünder. 2022. Sentiment analysis tools in software engineering: A systematic mapping study. *Information and Software Technology* 151 (2022), 107018. https://doi.org/10.1016/j.infsof.2022.107018
[54] José Palazzo Moreira de Oliveira. 2003. Sistemas de informação e sociedade. *Ciência e Cultura* 55, 2 (2003), 39–41.
[55] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology* 64 (2015), 1–18.
[56] Orestis Piskioulis, Katerina Tzafilkou, and Anastasios Economides. 2021. Emotion detection through smartphone's accelerometer and gyroscope sensors. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 130–137.
[57] Julio Reis, Pollyanna Gonçalves, Matheus Araújo, Adriano Pereira, and Fabricio Benevenuto. 2015. Uma Abordagem Multilíngue para Análise de Sentimentos. In *Anais do IV Brazilian Workshop on Social Network Analysis and Mining* (Recife). SBC, Porto Alegre, RS, Brasil, . https://doi.org/10.5753/brasnam.2015.6767
[58] Luis Rivero and Tayana Conte. 2017. A Systematic Mapping Study on Research Contributions on UX Evaluation Technologies. In *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems* (Joinville, Brazil) *(IHC 2017)*. Association for Computing Machinery, New York, NY, USA, Article 5, 10 pages. https://doi.org/10.1145/3160504.3160512
[59] Cecilio Merlotti Rodas, Suellen Elise Timm Barros, Rodrigo Ananias da Silva Souza, and Silvana Aparecida Borsetti Gregorio Vidotti. 2022. Análise de sentimentos sobre as vacinas contra Covid-19: um estudo com algoritmo de machine learning em postagens no twitterr. (2022).
[60] Stuart J Russell and Peter Norvig. 2010. *Artificial intelligence: a modern approach*. Pearson Education Limited.
[61] J Alfredo Sánchez, Ximena Cortés, Oleg Starostenko, Ofelia Cervantes, and Wanggen Wan. 2015. An extensible platform for seamless integration and management of applications for emotion sensing and interpretation. *Journal of Ambient Intelligence and Smart Environments* 7, 1 (2015), 5–19.
[62] Rosario Sanchis-Font, Maria Jose Castro-Bleda, José-Ángel González, Ferran Pla, and Lluís-F Hurtado. 2021. Cross-Domain polarity models to evaluate user eXperience in E-learning. *Neural processing letters* 53, 5 (2021), 3199–3215.
[63] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing Management* 45 (07 2009), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
[64] Meishu Song, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Zijiang Yang, Shuo Liu, Zhao Ren, Ziping Zhao, and Björn W Schuller. 2021. Frustration recognition from speech during game interaction using wide residual networks. *Virtual Reality & Intelligent Hardware* 3, 1 (2021), 76–86.
[65] Ehsan Jahangirzadeh Soure, Emily Kuang, Mingming Fan, and Jian Zhao. 2021. CoUX: collaborative visual analysis of think-aloud usability test videos for digital interfaces. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 643–653.
[66] Arnold POS Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. 2010. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries*. 521–530.
[67] Maria Virvou. 2023. Artificial Intelligence and User Experience in reciprocity: Contributions and state of the art. *Intelligent Decision Technologies* 17, 1 (2023), 73–125.
[68] Tanja Walsh, Jari Varsaluoma, Sari Kujala, Piia Nurkka, Helen Petrie, and Chris Power. 2014. Axe UX: Exploring long-term user experience with iScale and AttrakDiff. In *Proceedings of the 18th international academic mindtrek conference: Media business, management, content & services*. 32–39.
[69] Jiaqi Zhang, Eiji Kamioka, and Phan Xuan Tan. 2019. Emotions detection of user experience (Ux) for mobile augmented reality (mar) applications. *International Journal of Advanced Trends in Computer Science and Engineering* 8, 1.4 S1 (2019), 63–67.